Welcome To

UNIVERSITI TEKNOLOGI MALAYSIA

The Discovery University

Mining Hidden Information From Library Databases Using SOM (Self Organizing Map)

By

PM. DR. Mohd Noor Md Sap

Sarjon Defit

University of Technology Malaysia (UTM), Skudai

University of Putera Indonesia (UPI) Padang, West Sumatera

Agenda

- Introduction
- Mining Hidden Information (MHI) Model
- Testing and Experimental Results
- Conclusion
- Question Answer

Introduction

In era the Internet and Distributed of Information System Applications, the Proliferation of

- **Textual and Multi Media Database**
- **Digital Libraries**
- **Internet Servers'**
- **Intranet Services**

Has Increased Rapidly

Introduction

It Has Turned Researcher's and Practitioners' Dream of Creating an Information Rich Society Into a Nightmare of Information Gluts.

Turning an Information Glut Into a Useful Digital Library Requires a Powerful Methods for Organizing, Exploring, and Searching Collection of Free Form Textual Documents.

Previous Researchs

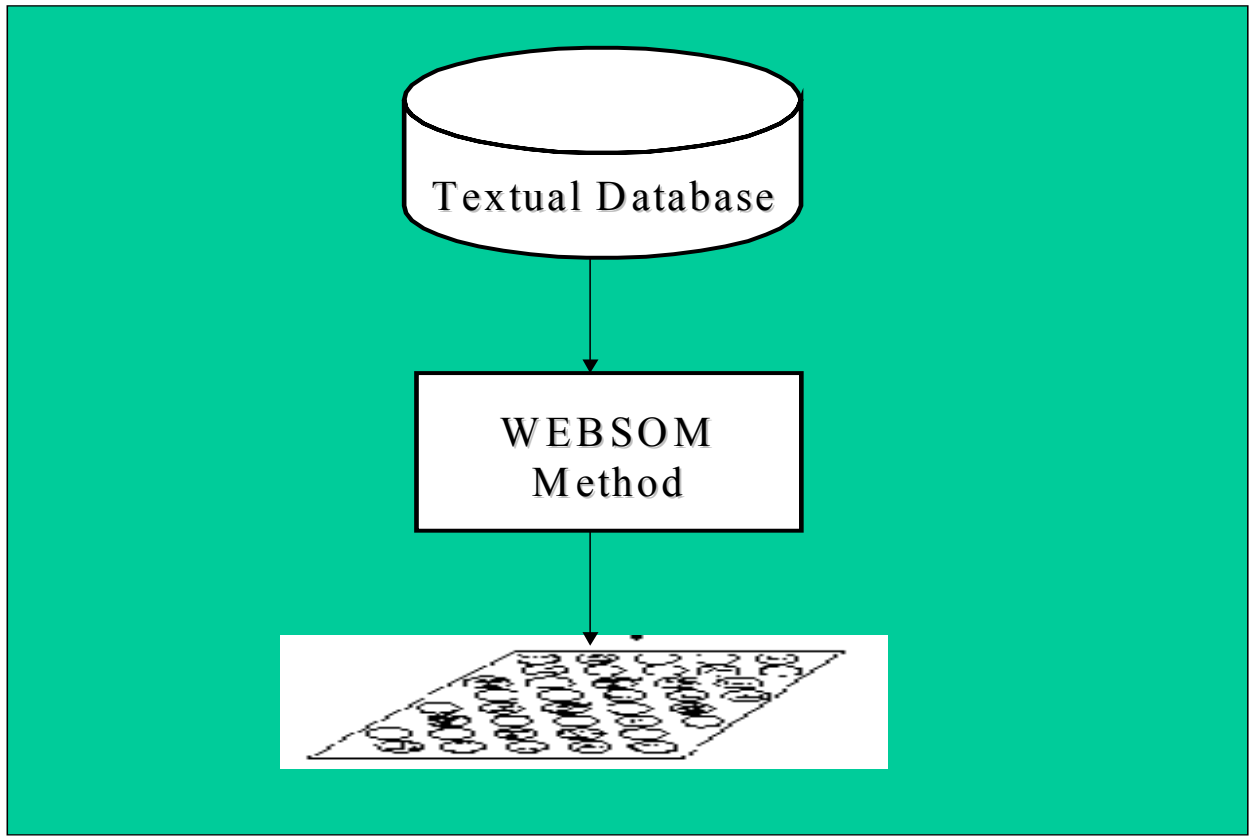
Researchers	Method
Kaski, S., Honkela, T. et.al 1996	An Explorative Full Text Information Retrieval Method Based on SOM (Self Organizing Map) Algorithm to Order Documents Based on Their Full Text Contents
Roussinov, D.G., Chen, H., 1998	A Scalable Textual Classification and Categorization System Based on the Kohonen's Self Organizing Feature Map (SOM) Algorithm

The MHI Model

A Mining Hidden Information (MHI) Model From Textual Database Using WEBSOM to Organizes a Document Collection on Map Display that Provides an Overview of the Collection and Facilitates Interactive Browsing

The General Architecture of MHI Model :

The MHI Model



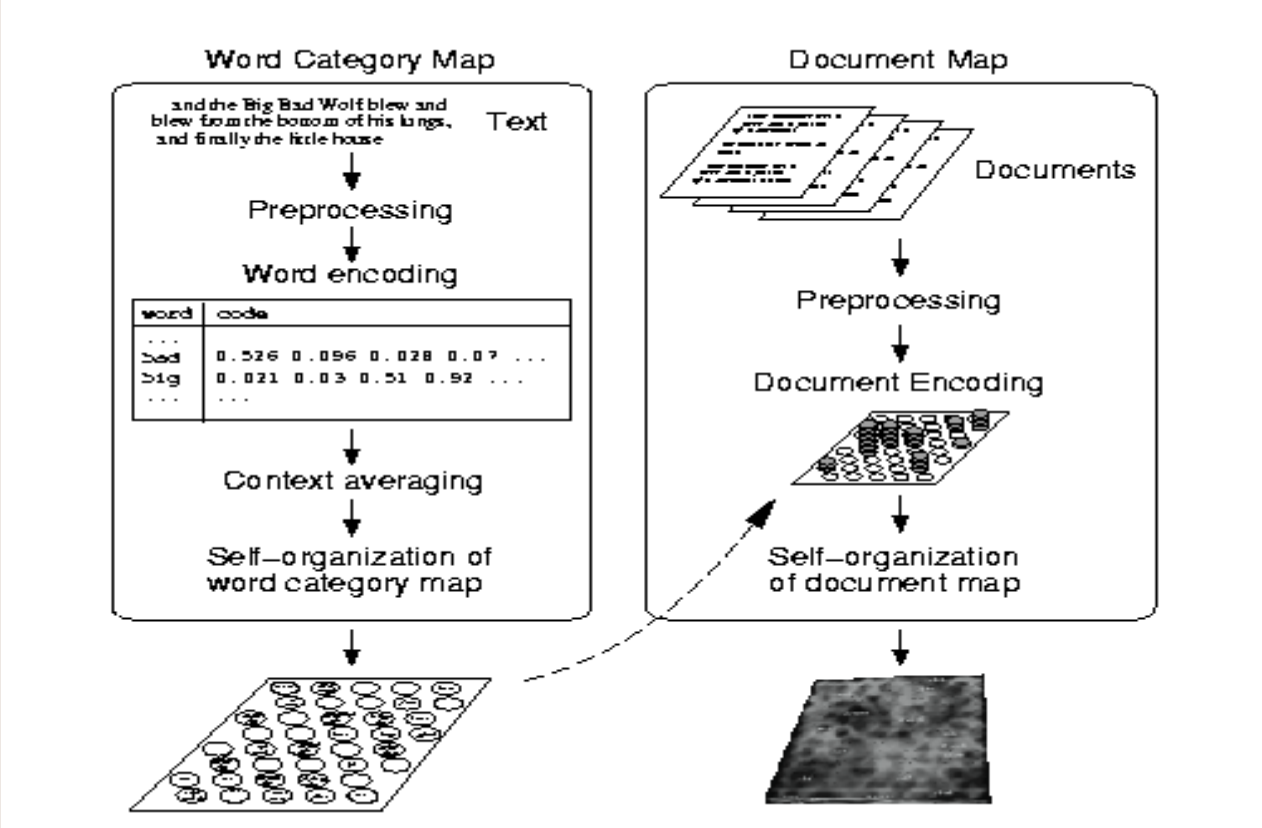
WEBSOM

WEBSOM is a Method for Organizing Miscellaneous Text Documents Onto Meaningful Maps for Exploration and Search.

It based on the SOM algorithm that Automatically Organizes the Documents Onto a Two Dimensional Grid so That related Documents Appear Close to Each Other

WEBSOM

The Overall Architecture of the WEBSOM Model



WEBSOM

It Consists of Two Levels : The Word Category Map and The Document Map

The Document Map is Organized on Documents Encoded With the Word Category Map

Both Maps Are Produced With the SOM. When the Maps Have Been Constructed, the Processing of New Documents is Much Faster.

The Main Phase Include : Preprocessing of the Input, Formation of the Word Category Map, and Formation of the Document Map.

Testing and Experimental Results

Using Collection of Usenet Newsgroups Articles / Documents.

From June 1995 to March 1997

It Consists of 32627 Articles Containing a Total of Approximately 8511391 Words

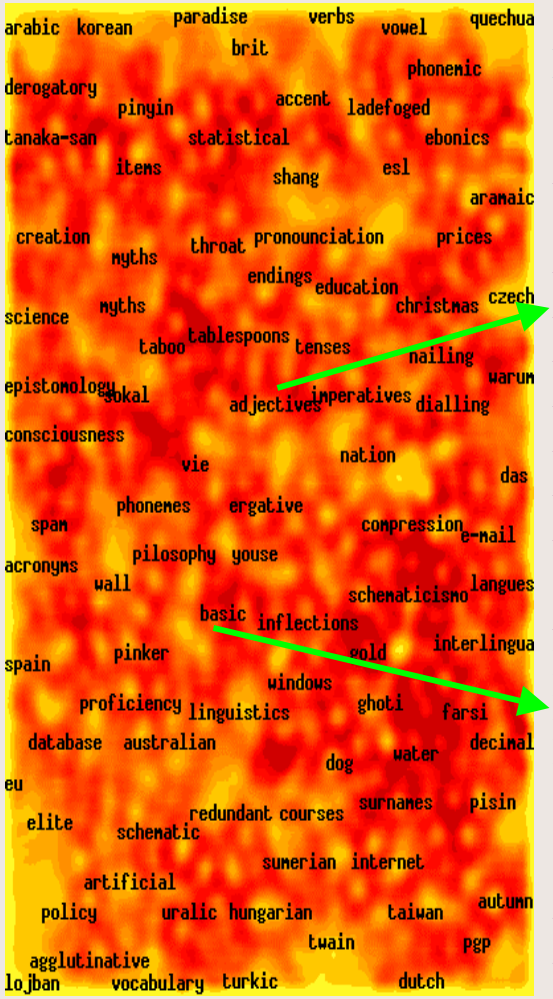
For Instance, “Number With gender”, Published on Sunday 27 Augustust 1995, is One of Article Used in Our Study.

Testing and Experimental Results

The WEBSOM Browsing Interface is Implemented as a Set of HTML Documents That Can Be Viewed Using a Graphical WWW Browser

The Whole or WEBSOM Map From “Number of Gender” Article

Testing and Experimental Results



Automatically Generated Labels and Examples of Titles in which the Labels have Occured

Accent - German/Swiss Accent in English, was Re: Lowlands languagelist
Acronyms - Acronyms?
Adjectives – question on adjectives
Agglutinative – Inflected versus agglutinative languages
Arabic – intensive summer Arabic program in Alexandria, Egypt
Aramaie – Define Aramaie/Syriac boundary??
Artificial – alt. Language.artificial
Australian – ANNOUNCE: Australian Speech Science and Technology Association URL
Autumn – Lat CfP: Autumn School of GLDV, Sept 23-27, Magdeburg, Germany
Basic – Basic English
Birt – Brit vs Amer SIMPLE QUESTION
Christmas – Merry Christmas
…
…
Windows – Changing Alphanumeric Sort order in Windows
Youse – You, Youse, … All Y'all

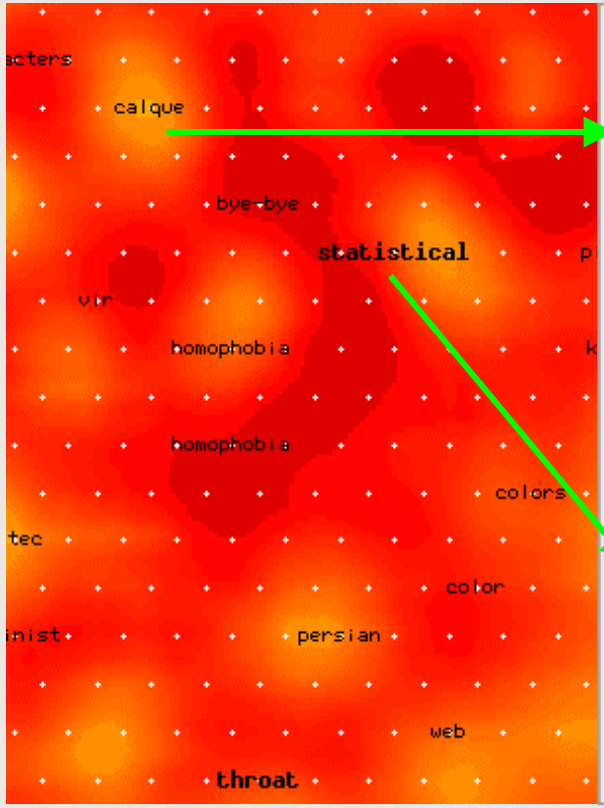
Testing and Experimental Results

The Left Side of This Figure Shows the WEBSOM or Whole Map, and the Automatically Generated Labels and Examples of Tittles in Which the Labels Have Occurred in the Right Side

For Instance, "Database" Label is Generated From :"IPA: Speech Database With Example Available?".

Testing and Experimental Results

A Zoomed Map From This Figure :



aztec - Maya vs. Aztec vs. Inca
bye-bye - Word doubling (e.g. bye-bye)
calque - The French word "calque" for loan translations
characters - Chinese characters vs. Latin(Roman)isation
color - Colors (was: Dialects (Was Re: Shakespeare's Future))
colors - Colors (was: Dialects (Was Re: Shakespeare's Future))
feminist - Lesbian feminists? (was: same old)
homophobia - the word homophobia
persian - Persian etymology
statistical - Statistical linguistics figures
throat - "Deep Throat"
vir - Sic Transit Vir
web - Grammatical gender of the Web

Testing and Experimental Results

Where:

Each White Dot Marks a Map Node.

Color Denote the Density or the Clustering Tendency of the Documents

White Areas are Clusters, and

Dark Areas Empty Sparse Between the Clusters.

Testing and Experimental Results

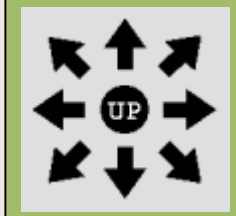
The Left Side Shows the Zoomed View, and

The Automatically Generated Labels and Examples of Titles in Which the label Occur in the Right Side

For Instance, “Statistical” Label is Generated From “Statistical Linguistics Figures” Document.

Testing and Experimental Results

List of Usenet Newsgroups Articles or Map Node From Zoomed Map :



Re: numbers with gender ,Joseph C Fineman, Sun, 27 Aug 1995, Lines: 22.

Statistical linguistics figures , Franck Noël, Fri, 17 Jan 1997, Lines: 14.

Testing and Experimental Results

This Figure Shows That " Statistical" Label is Generated From 2 Articles, namely :

a. **"Re: Number With Gender", Published on 27 Aug 1995**

b. **"Statistical Linguistic Figures", Published 17 Jan 1997**

Testing and Experimental Results

The Content of " Statistical Linguistic Figures"

220 66042 <32DF47A0.3D86@hp.com> article
From: Franck Noël <franck_noel@hp.com>
Newsgroups: sci.lang
Subject: Statistical linguistics figures
Date: Fri, 17 Jan 1997 10:34:24 +0100
X-Mailer: Mozilla 3.0 (WinNT; I)

Hello,

I'm currently writing a 'KeyWord Extractor' which is a tool that proposes a list of relevant words after scanning a document.

I would like to have some statistical figures (if it exists) such as : a word which appears XX% times in a document is probably useless or important or whatever.

Conclusion

The WEBSOM is Readily Applicable to Any Kind of Collection Textual Documents.

It is Especially Suitable For Exploration Tasks in Which The Users Either Do Not Know the Domain Very Well, or They Have Only a Limited Idea of the Contents of the Full Text Database Being Examined.

With the WEBSOM, the Documents are Ordered Meaningfully According to Their Contents.

Map Also Help the Exploration by Giving an Overall View of What the Information Space Looks Like

Question Answer

Thank You